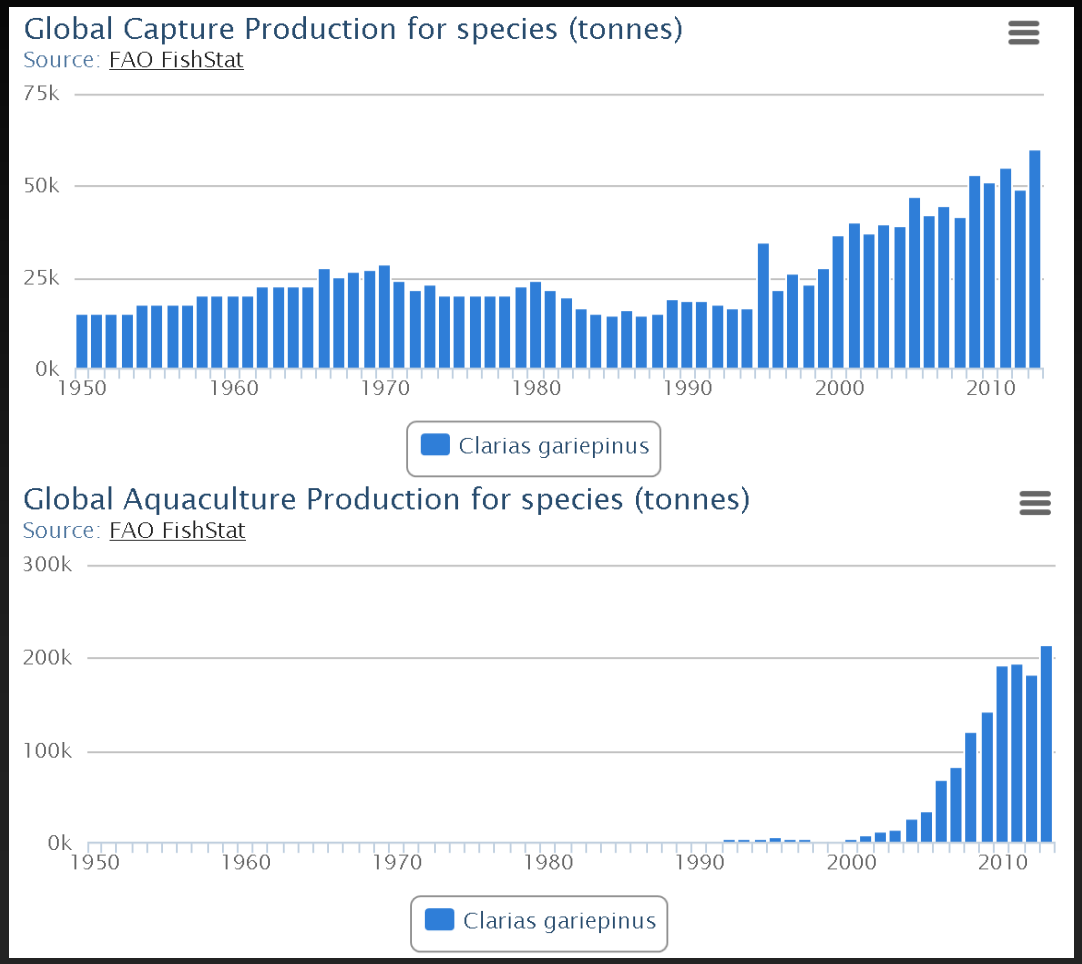
DE NOVO GENOME ASSEMBLY OF THE AFRICAN CATFISH (*CLARIAS GARIEPINUS*)

Kovács B.[a, §], Barta E. [c], Pongor S. L. [b], Uri Cs. [a], Patócs A. [b], Orbán L. [d], Müller T. [a], Urbányi B. [a]

[a] *Department of Aquaculture, Szent István University, Hungary*
[b] *Molecular Medicine Research Group, Hungarian Academy of Sciences and Semmelweis University, Hungary.*
[c] *Agricultural Genomics and Bioinformatics Group, National Agricultural Research and Innovation Center, Hungary.*
[d] *Reproductive Genomics Group, Temasek Life Sciences Laboratory, Singapore*

African catfish



- Important food fish
- Growing production
- The second on the list of cultured fish species in Hungary, following the carp (~ 2,000 tonnes)

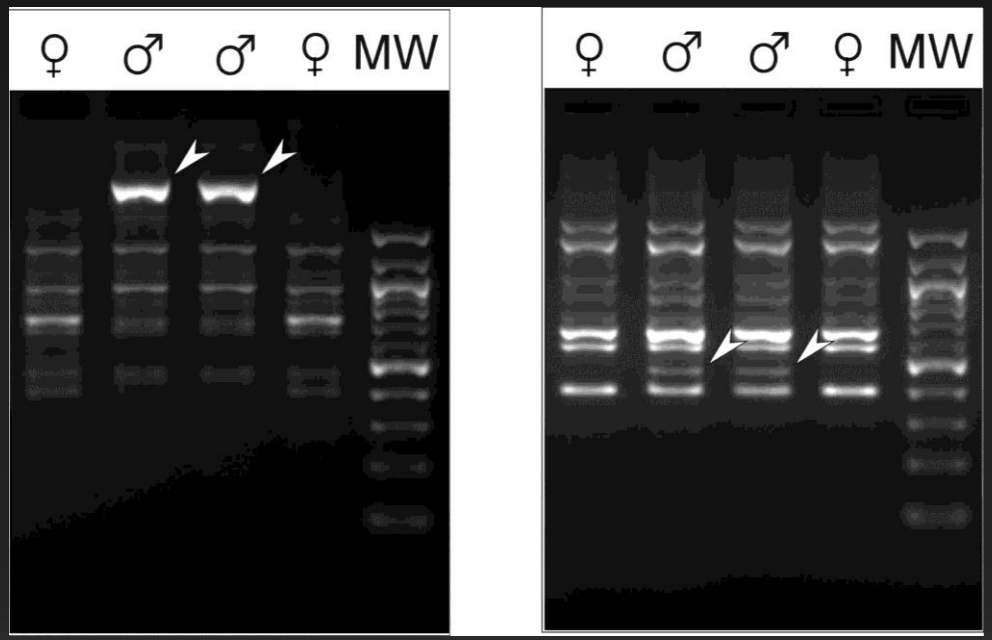
Main producer countries of *Clarias gariepinus*

(FAO Fishery Statistics, 2006)

More than 29 countries

Sex-specific DNA markers

- RAPD (more than 300 primer combination)
- TWO male-specific markers (ClgY1; ClgY2)
- XX/XY sex chromosomal system

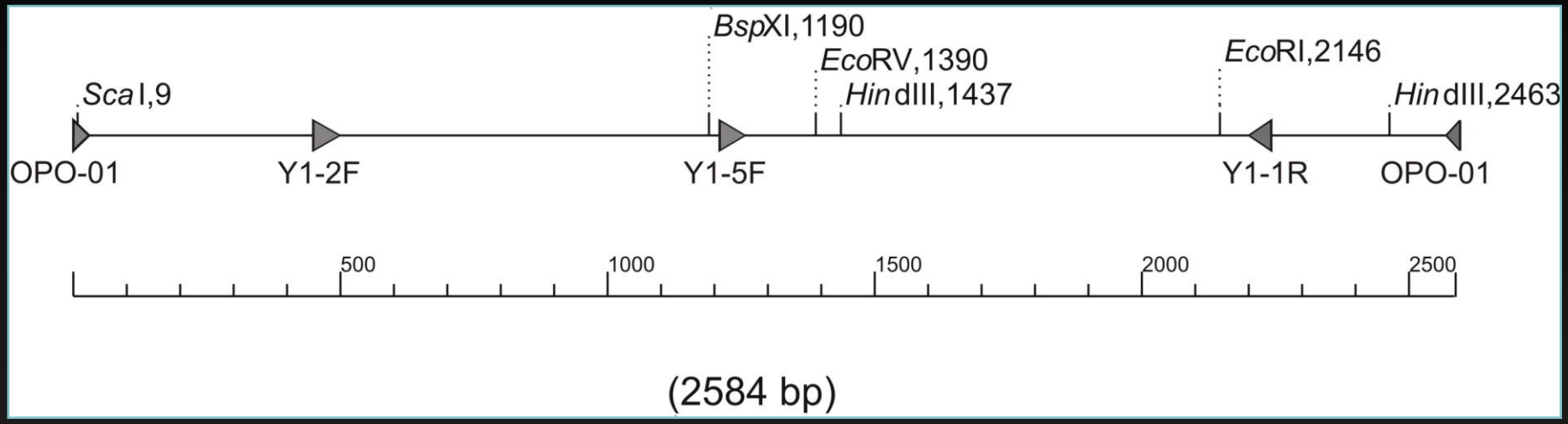


Kovács et al., Genetica (2001)

ClgY1

AT/(AT+GC) = 61%

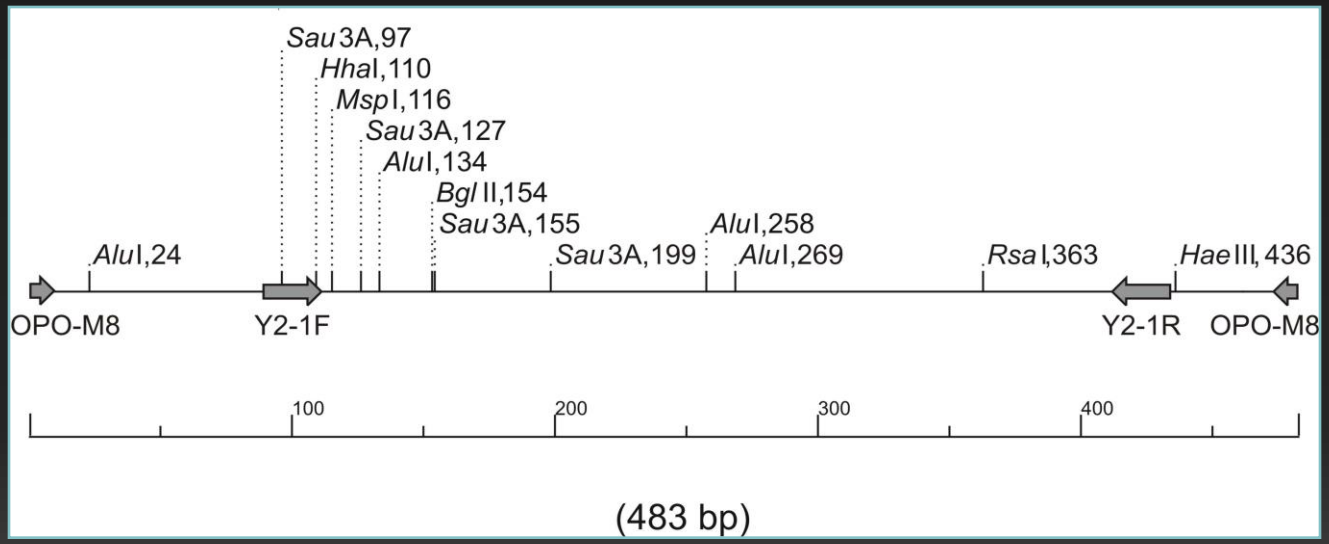
GenBank: AF332597

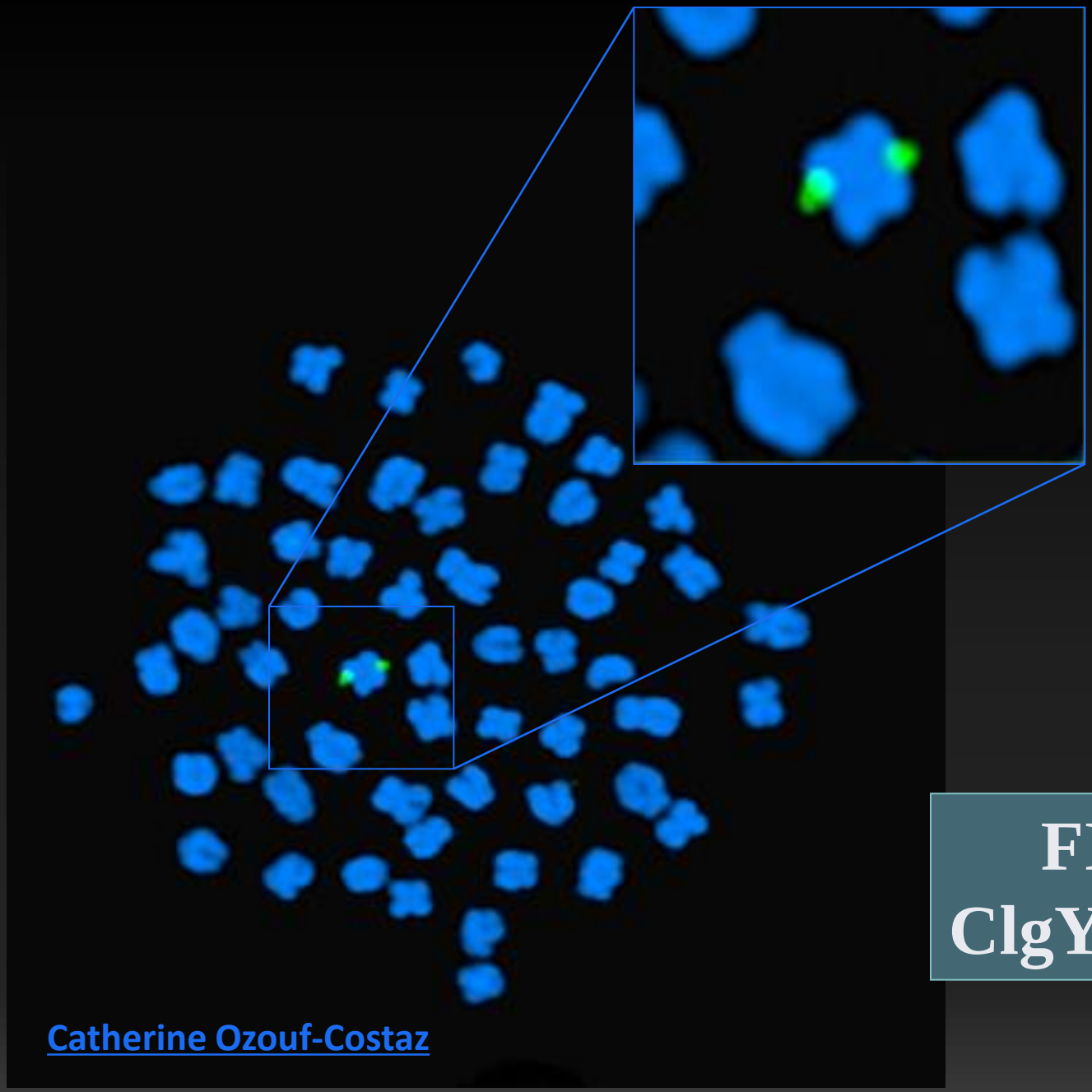


ClgY2

AT/(AT+GC) = 61%

GenBank:
AF332598





FISH
ClgY1 probe

Catherine Ozouf-Costaz

Test cross with separated rearing

	CglY1 test	phenotype	difference	Efficiency
Male	87	82	5	94,2 %
Female	92	91	1	98,9 %
Total	179	173	6	96,6 %

Distance between the marker and the gene ~3,4 cM
(1 cM corresponds to 500-600 kb DNA in the zebrafish)

Introduction

The available DNA-RNA sequence information and data on genetic background of the species is very limited.

- Microsatellite markers
- Mitochondrial fragments
- Sex-specific markers

Genomics information

- 2n=56
- C - value: 1.20pg

  (Animal Genome Size Database)
- Predicted genome size:

  1,173Mb

(Okonkwo and Obiakor; 2010)

Genome sequencing

Illumina Hiseq 2000

	MatePair	Paired-End	Total
Insert size (bp)	5,000	350-550	
Number of reads	430,208,098	281,447,634	711,655,732
Read length (bp)	100	100	
Coverage	37X	24X	61X

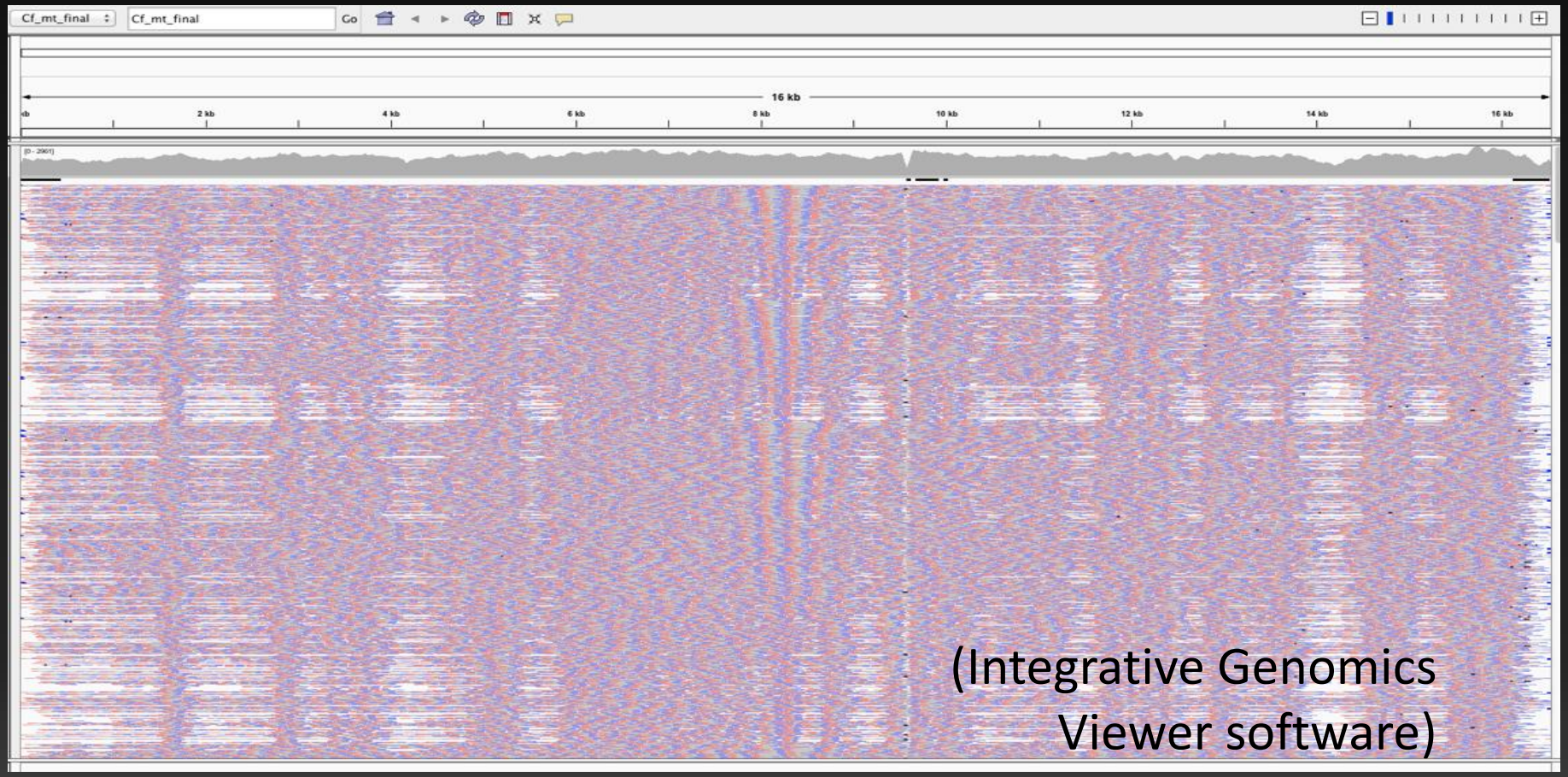
Mitochondrial genome assembly

- 2x100bp PE reads
- *Mirabait* (MIRA 4.0.2 package) - select mitochondrial reads
  - baiting – *Cl. fuscus* MT (reference) genome (KM029965)
- Mitobait.pl - to build the MT genome

– the ends were trimmed by hand

Assembled MT reads

- The full length: 16 543 bp
- A gap with low coverage (ca. 9.5 kb)→ confirmed by Sanger sequencing



Annotation of MT genome

MITOS- software

- 37 gene

Name	Start	Stop	Length	Strand
tRNA^{Phe}	1	69	69	H
12S rRNA	70	1021	952	H
tRNA^{Val}	1022	1093	72	H
16S rRNA	1094	2769	1676	H
tRNA^{Leu}	2768	2842	75	H
ND1	2846	3811	966	H
tRNA^{Ile}	3823	3894	72	H
tRNA^{Gln}	3894	3964	71	L
tRNA^{Met}	3964	4033	70	H
ND2	4034	5071	1038	H
tRNA^{Trp}	5079	5149	71	H
tRNA^{Ala}	5153	5221	69	L
tRNA^{Asn}	5223	5295	73	L
tRNA^{Cys}	5327	5393	67	L
tRNA^{Tyr}	5405	5474	70	L
COX1	5482	7017	1536	H

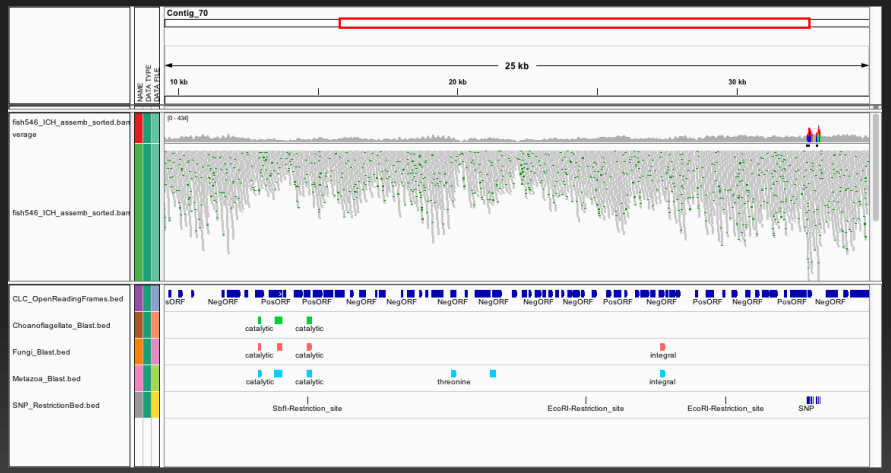
Name	Start	Stop	Length	Strand
tRNA^{Ser}	7027	7097	71	L
tRNA^{Asp}	7102	7174	73	H
COX2	7189	7872	684	H
tRNA^{Lys}	7880	7953	74	H
ATP8	7955	8119	165	H
ATP6	8113	8793	681	H
COX3	8796	9578	783	H
tRNA^{Gly}	9609	9681	73	H
ND3	9682	10029	348	H
tRNA^{Arg}	10031	10099	69	H
NAD4L	10100	10393	294	H
NAD4	10390	11763	1374	H
tRNA^{His}	11771	11840	70	H
tRNA^{Ser}	11841	11906	66	H
tRNA^{Leu}	11909	11981	73	H
NAD5	11991	13796	1806	H
NAD6	13808	14323	516	L
tRNA^{Glu}	14324	14392	69	L
Cyt b	14394	15521	1128	H
tRNA^{Thr}	15532	15604	73	H
tRNA^{Pro}	15603	15672	70	L

Nuclear Genomic Data processing

- National Information Infrastructure Development (NIIF) Program

  - (SGI UltraViolet 1000 – 1152 core 6 terabyte memory)

- *Quality selection: Trimmomatic*
  - 89 % of the data are usable
- De Novo Genome Assembly:
  - *SOAPdenovo*
  - *Allpaths-lg*
  - *Minia*



Results of genome assembly

	SOAPdenovo	Allpaths-lg	Minia
contigs	1,825,968	12,766	2,652,684
Largest contig	953,317	545,339	163,814
Total length	1,162,594,395	939,385,245	698,425,129
GC (%)	38.37	38,23	37.99
N50	49,277	600,219	10,176
#Ns per 100 kb	28,709	26,954	19,473
#predicted genes (unique)	155,876	96,820	117,608

Annotaton
Clariidae Siluridei UniProt sequences

	Clariidae family	*Siluridei* suborder
No of all genes	722	15,808
Identified genes	518	11,867
%	72	75
Percentage of genes showing similarity over 80%	55%	46%

Sex markers in the sequence

- Both marker were identified
  - ClgY1 (scaffold_4745, Length= 10,667)
  - ClgY2 (scaffold_1541, Length = 107,434)

Summary

- *The Cl. gariepinus MT genome is annotated*
- We have produced a draft sequence the *Cl. g. genome*
- The annotation of the genes is in progress
- Two scaffolds containing the sex markers have been identified
- Additional genome sequencing are planed
  - different insert size
  - longer reads
  - transcriptome

Acknowledgements

This work was supported by OTKA (105393) project, the National Information Infrastructure Development (NIIF) program and the Ministry of Human Resources of Hungary (contract number 8526-5/2014/TUDPOL).

Thank you for your attention!